University of Zagreb

Faculty of Humanities and Social Sciences

Department of English

TEFL Section

PERCEPTION OF COGNATE SIMILARITY

Graduation Thesis

Student: Nives Kovačić

Supervisor: Dr. Stela Letica Krevelj, postdoc

Zagreb, 2019

Sveučilište u Zagrebu

Filozofski fakultet

Odsjek za anglistiku

Katedra za metodiku

PERCEPCIJA SLIČNOSTI KOGNATA

Diplomski rad

Student: Nives Kovačić

Mentor: dr.sc. Stela Letica Krevelj, poslijedoktorandica

Zagreb, 2019.

Examining Committee:

Asst. Prof. Renata Geld

Jasenka Čengić, assistant

Dr. Stela Letica Krevelj, postdoc

Contents

1. Introduction 1
2. Previous research 2
2.1. Item-related variables 4
2.2. Participant-related variables 7
3. Study 8
3.1. The aim and research questions 8
3.2. Participants 8
3.3. Method 9
4. Results and discussion 11
4.1. Hungarian 11
4.2. Czech 14
4.3. Spanish 16
4.4. French 19
4.5. German 22
4.6. Swedish 24
4.7. Psychotypology at the language system level 26
5. Limitations 29
6. Conclusion 30
Appendix 35

Abstract

It has been widely acknowledged that language typology that is, the structural distance between languages which can be objectively measured, plays an important role in the process of language acquisition and processing. However, as it is often the case with any matter that involves human cognition and perception, humans do not perceive the objective distance between languages in the same way. About four decades ago Kellerman (1983) introduced the term psychotypology to refer to the perception of the degree of typological proximity which strongly influences the extent to which one will attempt to transfer from one language to another. The perception of similarity at the lexical level is the focus of the present study and it is our aim to shed some light on the variables which can influence it. An instrument consisting of 54 lexical items in Croatian, paired with their cognates in six different languages (nine cognate pairs per language), was given to 110 Croatian participants in order to measure their perception of similarity between the cognates. Based on the previous research into subjective similarity and receptive multilingualism, there are two types of variables which affect the subjective similarity: item-related and participant related. Their effect is tested in the study. In addition, the psychotypology at the language system level is believed to influence the similarity perception at the level of items as well. The results of the study generally corroborate some of the previous findings related to the item-related variables but there are also some unexpected findings in the similarity ratings accross six different languages. The study suggests that psychotopology at the language system level plays a role in similarity ratings of lexical items.

Key words: psychotypology, cognates, crosslinguistic similarity

Sažetak

Postoji generalni konsenzus kada je riječ o jezičnoj tipologiji, to jest, strukturnoj udaljenosti između jezika koju je moguće objektivno izmjeriti, i njenoj važnoj ulozi u procesu usvajanja i procesiranja jezika. Ipak, kako često i biva slučaj s pojavama koje uključuju ljudsku kogniciju i percepciju, ljudi ne percipiraju objektivnu jezičnu udaljenost na isti način. Prije otprilike četiri desetljeća, Kellerman (1983) uvodi termin psihotipologija, koji označava percepciju stupnja tipološke bliskosti koja snažno utječe na transfer elemenata iz jednoga jezika u drugi. Percepcija sličnosti na nivou leksika je fokus ovoga istraživanja, a cilj je pokušati pojasniti varijable koje mogu utjecati na percepciju. Upitnik s 54 riječi na hrvatskome jeziku te njihovim ekvivalentima u šest različitih jezika (devet parova kognata po jeziku) ispunilo je 110 hrvatskih ispitanika kako bi se izmjerila njihova percepcija sličnosti tih kognata. Temeljem prethodnih istraživanja subjektivne sličnosti i receptivne višejezičnosti, postoje dva tipa varijabli koje utječu na subjektivnu sličnost: varijable povezane s kognatima te sa samim ispitanicima. U ovom se istraživanju ispituje njihov efekt. Također, smatra se da psihotipologija na razini jezičnoga sustava također utječe na percepciju sličnosti na razini riječi. Rezultati istraživanja generalno potvrđuju neke od prethodnih pronalazaka, no došlo je i do neočekivanih rezultata u varijablama povezanima s kognatima u ocjenjivanju sličnosti šest jezika. Istraživanje ukazuje na to da psihotipologija na razini jezičnoga sustava ima ulogu u procjenjivanju sličnosti riječi.

Ključne riječi: psihotipologija, kognati, sličnost

1. Introduction

When asked about the languages they know, whether in a job interview or in a study in applied linguistics, most people first mention the language(s) they have learned and used in terms of language production. However, drawing on experience with language learners and users, one has encountered a number of people who emphasize and appreciate their receptive skills in a foreign language and consider the given language an important part of their repertoire. The level of comprehension seems to be essential to these language users.

What makes these individuals understand parts of an unknown language? As Ringbom (2006) explains, learning, including language learning, is based on prior knowledge. The role of the languages in the repertoire is of great importance, and the existence of crosslinguistic similarities has a largely facilitative effect in language learning. Ringbom (2006) emphasizes that "from the very beginning learners profit from similarities they perceive, especially formal similarities, which help them to establish cross-linguistic equivalences" (p.92). Therefore, language typology, that is, the structural distance between languages which can be objectively measured, plays an important role in the process of second language acquisition and processing. In other words, the closer the languages in the repertoire are to the target language, the easier it is to understand and learn it. However, as can be seen in Ringbom's quote, learners only profit from similarities they perceive, that is, the existence of objective, typological similarity does not guarantee transfer. Jarvis and Pavlenko (2008) maintain that subjective similarity affects how much the learner relies on the source language when learning or using the target language, whereas objective similarity can determine whether the transfer will be positive or negative.

It can be inferred that all humans do not perceive the objective distance between languages in the same way. About four decades ago Kellerman (1983) introduced the term psychotypology to refer to the perception of the degree of typological proximity which strongly influences the extent to which one will attempt to transfer from one language to another. Psychotypology can also be defined as the subjective judgement of the similarity between languages.

The studies on psychotypology are not plentiful; for example, Letica Krevelj (2014) examined the role of psychotypology when choosing the source for transfer in third language production, whereas Kaivapalu and Martin (2017) compared objective and perceived

similarity scores of Estonian and Finnish. The studies which look into the variables which affect receptive multilingualism, namely, cognate guessing and recognition in an unknown language, are also considered important for psychotypology since the perception of similarity underlies and precedes these processes. For example, a person who has Croatian as a first and English as a foreign language in their repertoire understands some Spanish. If that person did not perceive Spanish as similar to Croatian (or English) to some extent, the odds of them observing similarities would be significantly lower. What is more, if they did not judge the crosslinguistic similarities to be more or less similar to the structures in their repertoire, they would not be able to recognize or understand them.

It is believed that by examining in which structures and why the disparity between objective and subjective similarity occurs, along with the properties of the learners' repertoires which influence the said disparity, more insight can be gained into how language learners and users perceive language. In addition, it is important to understand how the factors causing the disparity change across different languages, be it typologically similar or very different. In other words, the perception of a language on a macro level affects the subjective judgement of crosslinguistic similarities in that language as well. The focus of the present study is to examine the factors which affect the perception of similarity and how they change across six different languages. The variables which are examined were found to influence reception of cognates in previous research. Therefore, in the theoretical part of this paper the previous studies will be summarized. Special emphasis will be placed on the variables which have been proven important in receptive multilingualism as they provide groundwork for the interpretation of the results of the present research.

2. Previous research

Letica Krevelj (2016) found that lexis plays an important role in assessing the similarity even between unrelated languages (p.200). Accordingly, the existence of cognates may enhance the process of learning an unknown, even unrelated, language if the learner is able to judge the communicative utility of the cognate forms (Otwinowska-Kasztelanic, 2011). Cognates are traditionally defined as word forms which have descended from a common parent word (Schmitt, 1997, as cited in Otwinowska-Kasztelanic, 2011). However, a broader perspective

on cognates is adopted. Otwinowska-Kasztelanic (2011) explains that such view includes words borrowed from one language to another or borrowed independently by some languages. The latter category can be termed internationalisms or international words since the loanword occurs in a number of languages. Cognates which are broadly defined as historically related word pairs, which entail not only shared inherited words but also shared loans (Kurschner, Gooskens and Van Bezooijen, 2008), allow for the inclusion of a larger variety of languages. In other words, if indirect borrowing is included, typologically distant languages can also have cognate pairs.

Language learners and users will make us of crosslinguistic similarities, including cognates, if they perceive the target language to be similar to the source language(s) and/or if they perceive the cognate in the target language to be similar to the cognate in the source language(s). The former phenomenon is called psychotypology at the language system level, whereas the latter is called psychotypology at the level of items (Letica Krevelj, 2014). There are numerous factors which shape the perception of a certain language. For example, Gooskens (2007) maintains that extralinguistic factors such as language contact, attitudes, instruction should not be neglected in the studies of intelligibility as they affect the perception of a language system. Letica Krevelj (2014) has not found the connection between the choice of the source language for transfer and the psychotypology at the language system level; however, the author suggests it should be studied with respect to typologically very distant languages such as Croatian, Hungarian and Japanese.

It is argued that the psychotypology at the language system level may influence the perception of cognate similarity as well. Vanhove and Berthele (2017) encapsulate the psychotypology at the level of cognates in the following way: “some cognates are easier to recognize than others, and some readers or listeners are better able to recognize cognates than others” (p.2). Since one cannot recognize a cognate without perceiving it as similar to some extent, the same applies to cognate similarity perception. Accordingly, there are two types of variables which affect the similarity judgements of cognates: item-related and participant-related. Both will be described in the following subsection.

2.1. Item-related variables

The largest group of factors which influence the perception of cognate similarity are the formal differences between cognates, that is, the objective properties of cognate pairs which can be measured.

It has been mentioned in the introduction that the term typology refers to the formal distance between languages which can be objectively measured. A means of measuring the distance between words is the Levenshtein algorithm. The basic premise of the algorithm is that words are processed as strings of letters. Kaivapalu and Martin (2017) claim that "alphabetic writing systems make us see written languages as strings of letters which form words and sentences" (p.150), which influences the perception of language. With its roots in dialectology, Levenshtein distance (LD) can be defined as a string-matching algorithm which measures the number of operations necessary to transform one string into another (Letica Krevelj, 2014). In other words, it is a total minimal operation cost or the total number of insertions, deletions and substitutions. Furthermore, Levenshtein distance is related to the total length of the word – a longer word has more potential for a higher LD due to the bigger number of elements which can be different. In order to optimize the scores, length-normalized distance measure can be employed. As was done in the study by Vanhove and Berthele (2015), the total operation cost is divided by the length of the longest possible least cost alignment. To exemplify the proceeding, the calculation of length-normalized LD of the Spanish cognate *corbata* and the Croatian equivalent *kravata* is shown in Table 1.

Table 1: An example of Levenshtein distance calculation

1	2	3	4	5	6	7	8
c	o	r		b	a	t	a
k		r	a	v	a	t	a
1	2		3	4			

As can be seen from the table, once the strings are aligned, there are two consonant substitutions, a vowel insertion and a vowel deletion necessary to transform *corbata* into *kravata.* The overall LD score (4) is divided by the alignment length (8), which equals 0.50.

According to Vanhove and Berthele (2015), Levenshtein distance alone does not account for the similarity perception. However, the similarity judgement is higher when the formal overlap is high, which implies the important role of the formal distance. The importance of formal distance in receptive multilingualism has been shown in various studies on receptive multilingualism, both in written and oral tasks (Vanhove & Berthele, 2015; Möller & Zeevaert, 2015; Kaivapalu & Martin, 2017; Beijering, Gooskens & Heeringa, 2008; Kürschner, Gooskens & Van Bezooijen, 2008).

Other than the overall formal distance, the importance of consonants in the process of cognate recognition has been shown in various studies (Berthele, 2011; Vanhove & Berthele, 2015), whereas there was no significant difference between vowel and consonant differences in the study by Möller and Zeevaert (2015). However, it is claimed that consonants carry more information and that they are less variable than vowels, which is why they function as reference points when it comes to intelligibility (Gooskens, Heeringa & Beijering, 2008, p.64). An interesting example used to illustrate the importance of consonants in word identification is given by Ashby and Maidment (2005, as cited in Gooskens, Heeringa & Beijering, 2008, p.64) – if all the vowels in *Mary has a little lamb* are replaced by [ɛ], the majority of people could still understand the sentence. However, if all the consonants are replaced with [d], the sentence is incomprehensible. Although they are phoneticians and the example concerns phonetic differences, it adequately shows how consonants affect the recognition of words. Consonantal distance can also be calculated by means of the Levenshtein algorithm; however, in this calculation only the insertions, deletions and substitutions of consonants are accounted for.

Furthermore, it is not only the number of changes between two cognates, but also their qualitative properties that affect the similarity ratings. In other words, the operations involved in the calculation of the Levenshtein distance do not necessarily carry the same weight in word processing. For example, in her study Gooskens (2007) assigns less cost to substitutions of a vowel by a vowel or of a consonant by a consonant (0.5 point) than to insertions, deletions or substitutions of a vowel by a consonant or of a consonant by a vowel (1 point). It must be noted that diacritics are given the cost of 0.25 point. In their word recognition study,

Möller and Zeevaert (2015) found that substitutions were preferred over insertions or deletions, that is, the participants had less difficulty identifying a cognate pair when the changes between the strings did not include insertions or deletions.

Other than being particularly sensitive to consonants, it is hypothesized that the learners are affected by changes in the word onset more than by the differences in other word parts. This tendency is shown in cognate guessing and recognition studies by Vanhove and Berthele (2015), Berthele (2011) and Möller and Zeevaert (2015). Word onset is defined as the part of the word up to and including the first consonant or consonant cluster. There are various reasons for the possible reliance of the participants on word beginnings. Broerse and Zwaan (1966) maintain that the importance of initial letters in word identification is based on the fact that word onset contains more information. Also, they claim that words are retrieved in a sequential pattern and the initial letters are the starting point, which is in line with the implied linearity of the string-matching algorithm used to calculate formal distance (Levenshtein distance). Another explanation is that the importance of onset stems from the general psychological rule which is termed the "principle of least effort" by Zipf (1949, as cited in Broerse and Zwaan, 1966, p.445). The application of Zipf's principle in language and, more specifically, in the study of cognate similarity perception can be the following: if the word beginning of a cognate in the target language is identical to the word beginning of a cognate in the source language, a great deal of uncertainty is removed from the subject as there are fewer possibilities for the word to end due to phonetic and morphological constraints. The Levenshtein algorithm can also be used to calculate word beginning formal distance.

The following item-related determinant which proved to affect the perception of similarity is the neighbourhood effect. Neighbours are defined as words which have similar form, which suggests that the neighbouring words compete for lexical activation (Kürschner, Gooskens & Van Bezooijen, 2008). The implication of the neighbourhood effect is the fact that shorter words have more words with similar form, which leads to them being perceived as less similar and, consequently, less transferable. In their study on intelligibility of Swedish words among danes, Kurschner et al. (2008) found a correlation of neighbourhood density and intelligibility.

The overall formal distance between the Croatian *kravata* and the Spanish *corbata* has been calculated on the graphemic level. The overall phonetic distance between the two words can be determined by comparing the strings of phonemes and calculating the minimal operation

cost necessary to transform one string of phoneme into another one, which is analogous to the graphemic distance. In her study from 2007, Gooskens found that phonetic distance correlated with intelligibility, whereas no significant correlation was found with lexical distance. However, even when the participants are given a written cognate guessing or recognition task, there is a possibility of them self-pronouncing the words based on assumed grapheme-phoneme correspondences. Berthele (2011) terms this phenomenon "imagined phonology". The participants who self-pronounce the unknown words form what Meissner (1997, in Peyer, Kaiser & Berthele, 2010) calls a "hypothetical construct" of the possible phonetic correspondences based on the input and on the knowledge of other languages. Language learners form their "imagned phonology" based on assumed grapheme-phoneme correspondences. The process in which participants engage in speculation about potential rules that might explain a phenomenon without other input and on the basis of other knowledge is called abduction (Berthele, 2011). The learners speculate about pronunciation rules of an unknown language without having learned the language.

2.2. Participant-related variables

Vanhove and Berthele (2017) claim that precise relationship between formal distance and perceived similarity varies from learner to learner. They assert that the effect of formal distance has been studied averaged over all the participants, but that not only factors concerning item properties affect the similarity perception, but also variables concerning participants. Thus, the effect of distance is influenced by the breadth of the participants' repertoires on the one hand, and by their ability to deal with abstract patterns in a flexible way on the other (Vanhove & Berthele, 2017, p.3). Berthele (2008, as cited in Vanhove & Berthele, 2017) explains that people with larger and richer repertoires have greater perceptual tolerance, that is, they are more flexible in dealing with language input which deviates from the languages in their repertoire. In his study, Berthele (2011) confirmed that the participants with a larger multilingual repertoire perform better in cognate guessing tasks. What is more, the participants with a higher proficiency in languages related to the target language have an advantage. In his study, Berthele (2011) also found that age, vocabulary learning ability and English proficiency influence the processing of cognates.

Otwinowska-Kasztelanic's study from 2011 corroborates Berthele's findings (2011). The author examined the differences in perceiving cognates between bilinguals and multilinguals.

One of the findings was that only multilingual learners proficient in several languages tended to notice cognates and make conscious use of them. Otwinowska-Kasztelanic (2011) emphasizes that multilinguals have had more experience with language learning and use, more chances to interact with the environment in different languages and enhanced metalinguistic knowledge and awareness compared to bilinguals.

3. Study

3.1.The aim and research questions

When faced with objectively similar structures in two different languages, people's perception of similarity varies. The causes of the variation can be connected to the properties of the structures or the participants and their linguistic repertoires. The aim of the study is to examine the subjective similarity judgements of cognates in unknown languages by discovering certain patterns or general tendencies in the similarity ratings with the help of different types of variables. The formal similarity between Croatian and the target languages varies and the choice of cognates was aimed at including different item-related variables.

Therefore, the research questions are the following:

1) Which item-related variables affect the similarity perception of cognates?
2) Do speakers with larger linguistic repertoires rate the cognates differently than those with smaller repertoires?
3) Does additional knowledge of language(s) which are related to the target language affect the rating of the target language?
4) Does the psychotypology at the language system level influence the similarity perception?

3.2.Participants

In total, there were 110 participants. Only eight participants did not list English under the languages they know at least to some extent and they were excluded from the study.

Therefore, the total number of participants is 102. All participants were of Croatian nationality, therefore, they all have knowledge of Croatian aside from English. 54.9% of the participants also claim to speak one of the Croatian dialects.

Concerning age and gender of the sample, they are distributed so that the results are holistic, that is, so that they reflect the general tendencies of the population. Therefore, there are 56 participants aged 18-29 and 46 participants over 30 years of age. Also, there are 53 females and 49 males. Aside from the Croatian and English base, the participants differ in the number of languages and in the languages they have in their repertoires, given that they listed the total of 30 different languages. However, the participants with the knowledge of the target language are excluded from the analysis of the ratings of the language in question, which excludes the role of the target language knowledge.

3.3.Method

The subjects were asked to provide their personal information together with information on their language repertoires. They listed all the languages in their repertoires and assessed their proficiency on a Likert scale from 1 to 5. It is believed that even a certain contact or experience with a language changes the way a new language is processed, which is why even low proficiency was not excluded.

The subjects were asked to rate the similarity of the cognate in the target language and its equivalent in Croatian by drawing a cross on a directed line segment. The foreign word was on the left part of the line and the Croatian translation on the right. It has already been mentioned that the premise of the graphemic Levenshtein distance is processing written language as strings of words and sentences. The alphabetic system imposes the processing from left to right. Therefore, the foreign word is noticed and processed first; the participants start with the unfamiliar and new information and finish with the familiar, making the unfamiliar part more salient. What is more, the directed segment line on which the participants place crosses is used instead of a Likert scale. The positions of crosses have been measured with a ruler. Utgof (2008) used a similar method in her study, claiming that it allows the participants to follow their intuition rather than focus on choosing a number on a Likert scale. The extreme points of the line were marked with A and B. At the beginning of each group of cognates there was a reminder indicating the meaning of the points: A=“completely identical“ and B=“completely the same“. In order to avoid confusion and

possibly different ratings of the first few cognate pairs, a few practice items were given at the beginning. It is also important to note that the subjects were given a set of nine cognate pairs per language. The order of the target languages was changed in order to eliminate the fatigue effect. Due to the heterogeneity of the participants and the high number of languages in this research, the results are analysed with the help of descriptive statistics and qualitative explanations.

The cognates were chosen with the use of etymology dictionaries. The criteria were based on item-related variables so that there are words with different overall, consonantal Levenshtein distance, changes in the onset and in other positions, consonantal and vowel changes distributed across six languages. All cognate pairs are nouns. Furthermore, it was possible to calculate the correlation between the cognate similarity ratings and the number of languages in the repertoires. Also, based on their language repertoires two groups of participants were extracted. All the participants have certain knowledge of Croatian and English, that is, of a Slavic and a Germanic language. The first group entails the participants who have Slavic and Germanic languages in the repertoire, not considering the number or the variety. The second group are the participants who, on top of the Germanic and Slavic base, also have one or more Romance languages in the repertoire. The ratings of the two groups are compared, and the possibly significant differences will be termed "the Romance language effect". It must be noted that the roles of other languages in the participants' repertoires are not systematically taken into account in this research; however, they are used to explain the unexpected ratings of certain cognate pairs. What is more, the role of context is eliminated since the research design is deliberately reductionist, that is, the test items are isolated cognate pairs.

Other than the choice of items, the choice of languages was also delibarate. The idea was to have different language families in the study. Hungarian was chosen as the most typologically distant language, Czech as a representative of the Slavic group, French and Spanish as Romance languages and German and Swedish as Germanic languages. At the beginning of each group of cognates, the language to which the cognates pertain was clearly indicated in large, bold font. It is argued that the participants' awareness of the language in question affected the ratings as well, which could be indication of the psychotypology at the language system level. Comparing the mean ratings of the cognates for each language is a highly questionable method for measuring the effect of the psychotypology at the language system level since the item-related variables are not equally distributed nor controlled for in all six languages. However, the loanword *region* might be a better way of approaching the issue. The

overall Levenshtein distance is identical for all versions of *region* in the target languages and the position of the changes is identical (final position). The mean rating for the loanword is calculated and compared. The assumptions of the author concerning psychotypology of the languages in the study are the following. Hungarian is the only Non-Indo-European language in the study and it is typologically distant from Croatian. Notwithstanding the structural distance of the Uralic language, it must be noted that Hungary borders Croatia in the north-east, which is why it has had a certain influence on the varieties spoken in that part. The participants from this area might also be more acquainted with it. Nevertheless, it is hypothesized that Hungarian at the level of language system will be perceived as the least similar, which will affect the ratings of the cognates. Czech is the only Slavic language and there is a possibility that the participants find it quite similar to Croatian. In addition, when it comes to Romance languages, the hypothesis is that Spanish will be rated as more similar to Croatian than French. The reason behind this assumption is the popularity of the soap operas in the Spanish language which have permeated the daily life of Croatian people since the appearance of the popular *Santa Barbara* in the 1990s. The soap operas have brought about the familiarization with Spanish. Furthermore, receptive multilingualism implies language reception. Spanish is easier to understand than French since it is a phonetic language with mostly straightforward phoneme-grapheme correspondences, simple accentuation and spelling rules. Finally, due to extensive historical contacts with German as well as the omnipresence and popularity of learning German for economic reasons, it was hypothesized that Swedish will be perceived as more different than German.

4. Results and discussion

4.1. Hungarian

In the case of Hungarian cognates, it is confirmed that the overall formal distance affects the ratings together with the neighbourhood effect. Also, it is assumed that in certain words the ratings were higher due the fact that the phonetic distance is lower than the lexical distance, which confirms the hypothesis that some participants self-pronounce the words. The most surprising phenomenon in Hungarian is the importance of vowels. In some cognate pairs mostly vowel changes take place, and they were judged to be less similar than expected since

the consonants are mostly equal (*ecet, vacsora*). To sum up, in the case of Hungarian, typologically the least similar language to Croatian, vowels do affect how the participants perceive cognates. Finally, there is a tendency of the participants with a higher number of languages to be more tolerant to the consonantal changes. However, the correlations are rather mild.

The highest ratings were assigned to the cognates *csizma* and *szoba*. Despite the fact that both words have consonantal changes which are located near the beginning of the words, that is, which belong to the first consonant cluster, the high similarity rating is not unexpected. The Levenshtein distance for both words is low (0.17 and 0.20 respectively). Moreover, the phonetic distance of the highest rated Hungarian cognate *csizma* is zero since the Hungarian consonant cluster /cs/ and the Croatian /č/ are phonetic equivalents. This finding is in accordance with research which confirmed the importance of phonetic distance for intelligibility (Kurschner, Gooskens, & Van Bezooijen, 2008; Beijering, Gooskens, & Heeringa, 2008). However, despite the high similarity rating, the variance of *szoba* is relatively high (6.28).

The final word with high similarity rating is *só*. The LD remains quite low, however, in relation to the Croatian equivalent *sol*, it requires a consonant insertion, which proved to be rated quite low in other cognates. The possible reason could be the dialect of the participants. 54.9% of the participants claim to be dialectal speakers and in some Croatian dialects the final consonant in *sol* is omitted. Other than the dialectal effect, the monosyllabic cognate is the shortest. As mentioned, longer words are better recognized than shorter words due to the neighbourhood effect because shorter words have more competing word forms that are very similar to the stimulus word (Kurschner, Gooskens & Van Bezooijen, 2008)., but in the context of the present study it can only be hypothesized that it might have lowered the similarity rating score to some extent.

The following word pair, *ecet* and *ocat,* scored below expected in the similarity rating. Even though the overall LD is among the highest, the consonantal LD is zero since there are only two vowel substitutions. Contrary to some research findings (Berthele, 2011; Gooskens, Heeringa, & Beijering, 2008), vowel variation proved to have a significant effect on similarity judgement. However, it is impossible to generalize on a small sample with such a big number of variables which could affect the ratings, especially considering the psychotypological effect of the language in question being Hungarian. Furthermore, the lowest rated cognate,

vacsora, has a very low consonantal LD. With only one consonant deletion and two vowel substitutions necessary to transform it into the Croatian equivalent, it was perceived as the least similar. It appears that vowels in a Hungarian, a very typologically distinct language, do carry a lot of information value.

The following three cognates have the highest variance. In *palacsinta* and *görcs* the consonant cluster /cs/ is present in different parts – in the middle of the word and in the end. The fact that in *csizma* the same cluster is present in the onset and yet it is rated much higher could appear surprising. However, there are other operations necessary to transform *palacsinta* and *görcs* into their Croatian counterparts. In the latter, a consonant substitution is necessary, whereas in the former a vowel needs to be transformed into a consonant, which might have a higher operation cost (Gooskens, 2007). The variance in the word pairs containing the consonant cluster /cs/ could potentially be attributed to the fact that some participants engaged in self-pronunciations of the given words whereas some remained at the grapheme level.

Another unexpected rating is the low similarity perception of the cognate *kóró*. The LD is rather low and there is only one consonant insertion in the word. It is believed that the neighbourhood effect contributed to the low rating as the word for peel or crust in Croatian is *kora*, whereas the correct counterpart is *korov*. There is a possibility that the participants were confused by the consonant insertions and deemed it less similar than the vowel substitution necessary for the alternative, but incorrect word pair. In Table 2 all the ratings together with the overall LD and consonantal LD can be seen.

Table 2: Hungarian cognates

	Mean	Variance	Overall LD	Consonant LD
hu_vacsora	4,11	6,04	0,43	0,14
hu_görcs	4,20	6,74	0,40	0,20
hu_kóró	4,92	6,48	0,20	0,20
hu_palacsinta	5,85	6,84	0,20	0,20
hu_ecet	6,03	5,82	0,50	0,00
hu_só	7,17	5,78	0,33	0,33
hu_régió	7,43	5,56	0,33	0,17
hu_szoba	7,50	6,28	0,20	0,20

hu_csizma	8,11	4,32	0,17	0,17

The correlation of the number of languages in the participants' repertoires and the Hungarian similarity ratings proved significant in *palacsinta*, the cognate with the highest variance (6.84). In this cognate pair only consonantal differences are present. It might be that the nature of the differences, that is, the cost of two consonantal operations along with the importance of consonants caused the disparity among the participants. Despite the fact that the correlation coefficient was not high (r= .239), it can be stated that the more languages the participants speak the more tolerance they have for consonantal changes in this particular cognate pair. Similar can be said for *vacsora* (r=.219), *görcs* (r=.195) and *kóró* (r=.209). Despite rather modest correlation coefficients, the participants reacted differently to the cognate pairs with mostly consonantal changes. Furthermore, the comparison between the groups with and without a Romance language in their repertoires respectively yields no significant results.

4.2. Czech

The overall formal distance is again confirmed to influence the ratings; the higher the overall distance, the lower the ratings. As opposed to Hungarian, the participants have more tolerance to vowel changes in the Czech cognates; however, the importance of consonants is proven. It is interesting that in the case of *důkaz,* the diacritic seemed to lower the rating even though diacritics were systematically disregarded in the study. The nature of operations is also important; the participants were the least sensitive to vowel-vowel and consonant-consonant substitutions. Furthermore, differences in the onset contributed to the lower ratings, yet the participants with larger linguistic repertoires tended to be slightly less sensitive to word beginnings.

The two cognates which were perceived as very similar are *ucho* and *ořech*. Both words contain the consonant cluster /ch/, but it is positioned differently in each word. It is probable that the higher Levenshtein distance between *ořech* and *orah* resulted in lower rating despite the fact that their onsets are identical.

The following word took the unexpected fourth place. Even though the only LD operation is a vowel substitution, the rating of *důkaz* was lower than expected. A possible reason could be

the change in the first syllable or the novelty of the ring diacritic (˚). The diacritics were systematically disregarded in Levenshtein distance calculation since one of the aims of the present study is to descriptively compare the graphemic LD and the similarity perception ratings. However, the novelty of the diacritic in *důkaz* possibly contributed to the lower rating.

Pepř and *papar* form a cognate pair with the highest difference between overall LD (0.40) and consonantal LD, which is zero. When compared with *ecet* and *ocat,* the analogous cognate in Hungarian, it is perceived as more similar. What is more, the Czech equivalent does not only entail a vowel substitution, but also a vowel insertion. According to Gooskens (2007) and Möller and Zeevaert (2015), the cost of insertions is higher than the cost of substitutions. Nevertheless, the participants exhibited greater tolerance towards vowel changes in Czech, a typologically more similar language.

Křížovka and *lízátko* both have lower similarity ratings and high variance. In both cognates the differences take place at word endings. However, it is to be expected that *křížovka* would be perceived as more similar since the overall and consonantal LD are lower. Also, the substitutions either take place in consonants or vowels, which has been proven to be facilitative in cognate intelligibility (Berthele, 2011; Vanhove & Berthele, 2015). *Lízátko* is a cognate with the highest variance (6.99) and overall LD (0.50). What is more, two out of four operations necessary to transform it into *lizalica* are consonant-vowel and vowel-consonant substitutions. The great variance shows that some participants were more sensitive to these changes.

Even though it has one of the lowest scores in overall LD out of all the cognates (0.20), *hluma* is perceived as fairly different from *gluma.* The issue at hand concerns other types of item-related variables which affect language perception, namely word onset difference and importance of consonants. The difference in the first consonant in the word affects the perception as it evidently carries great information value. This cognate pair proves the importance of a holistic approach to item-related variables since the number of string operations is not as important factor as the type and the position.

This effect is especially prominent in the by far the lowest rated cognate pair, *výjimka* and *iznimka.* With three consonantal string operations in the first consonant cluster, the cognate was perceived as very different from its Croatian counterpart (3.32). Moreover, there is a significant correlation between *výjimka* and *hluma* and the number of languages in the participants' repertoire (r=.257 and r=.237), which implies a tendency of the participants with

less languages in the repertoire to be more sensitive to different onsets. All the data concerning the Czech cognates can be seen in Table 3.

Table 3: Czech cognates

	Mean	Variance	Overall LD	Consonantal LD
cz_výjimka	3,32	6,08	0,43	0,43
cz_hluma	5,24	6,82	0,20	0,20
cz_lízátko	5,82	6,99	0,50	0,38
cz_křížovka	6,46	6,11	0,33	0,22
cz_pepř	6,69	5,57	0,40	0,00
cz_důkaz	6,71	5,82	0,20	0,00
cz_ořech	7,18	5,58	0,40	0,20
cz_ucho	7,49	5,84	0,25	0,25
cz_region	8,07	4,22	0,33	0,33

4.3. Spanish

The ratings of the Spanish cognates confirm the influence of the overall formal distance and the word beginning on subjective similarity. It is also interesting that consonantal changes contribute to lower ratings. However, the participants with more languages in their repertoire tend to exhibit greater tolerance towards consonantal changes. Also, it has been observed that the participants make assumed grapheme-phoneme correspondences, that is, they rely on self-pronunciation when rating cognate similarity. In the case of *cocoa* and *kakao*, it is theorized that English was the supplier language given that all participants have English in their repertoires. What is more, the Romance language effect was not confirmed, which implies

that the participants did not rely on Romance languages in their repertoires when rating the similarity of Spanish cognates; yet it is plausible that English was the supplier language in *cocoa.*

The cognate pair that was perceived as highly similar is *pistola* and *pištolj*. The high result is in accordance with the premise of the research – overall and consonantal LD is low, the changes do not occur in the word onset and the existence of the diacritic is disregarded.

The word with the highest overall LD is *cocoa*. In order to transform it into *kakao*, three vowel substitutions and two consonant substitutions need to take place. According to the Levenshtein algorithm, *cocoa* and *kakao* are completely different strings. Still, the mean similarity rating is rather high (7.40). This cognate pair further displays the necessity of a more holistic and qualitative approach to explain language distance. There are several other variables which could provide explanations for the high similarity perception. Firstly, even though the overall LD is 1.00, the consonantal LD is 0.40. According to the abovementioned findings (Berthele, 2011; Gooskens, Heeringa, & Beijering, 2008), consonants do contribute more to word intelligibility and, analogically, to similarity judgement. Secondly, all participants have English in their repertoires, and the spelling of Spanish *cocoa* is the same as English. Due to the large quantity of heterogeneous data, Levenshtein distances were only calculated for the participants' L1 in this study. Berthele (2011), Berthele and Vanhove (2013) claim that cognate guessing is modelled more accurately when taking into account the possibility that the participants make use of multiple supplier Ls which is why they calculated the LD with respect to the participants' L2 and L3, which has not been done in the present study. However, in the case of *cocoa*, it is likely that the participants activated the English equivalent, which in turn affected the rating.

The next highly-rated cognate is *bicicleta.* Again, despite being the word with one of the biggest overall LD, it was deemed to be rather similar to the Croatian *bicikl.* The item-related variable which is recognized in this example is the importance of word onset, which is completely identical. At the end of the Spanish cognate, two vowel deletions, a consonant substitution and a consonant insertion are necessary to transform it into *bicikl.* Despite the large number of different types of operations, the cognate is perceived as very similar to the Croatian counterpart possibly due to the fact that the changes occur only at word ending. The opposite tendency takes place in the final highly-rated cognate pair, *cañón* and *kanjon.* There are two consonantal operations in the word onset necessary, yet the cognate pair was judged

as very similar (7.05). However, the high rating could be attributed to the phonetic LD, which is zero. This explanation would suggest that the participants correctly identified the correspondence between the Spanish /ñ/ and the Croatian /nj/ without the knowledge of Spanish language.

The following cognate pair is *huracán* and *uragan*. The mean similarity rating is 6.64, which is a bit higher than expected given that the first string operation is the deletion of a consonant. Nonetheless, the overall LD of *huracán* is 0.29, which is not a great difference. On the other hand, it is surprising that *huracán* is perceived as much more similar than *razón* (5.05). The onset in the cognate pair *razón* and *razum* is identical, whereas two substitutions occur at the word ending. However, in this case, the higher LD probably outweighed the position of the changes.

The second lowest rated cognate is *ojo*, the counterpart of the Croatian *oko.* The overall and consonantal LDs are relatively low and there is only one substitution necessary to transform one string into another. Also, in Czech the similar cognate is *ucho*, which has the second highest rating. Both the Spanish /j/ and the Czech /ch/ correspond to the Croatian /h/. It has come to light that the consonant deletion (/ch/ - /h/) is easier to the participants than the consonant substitution (/j/-/h/). What is more, *ojo* has the highest variance (7.79) and it is the only cognate with a significant correlation with the number of languages, r= .268. Again, there is a tendency of the participants with the higher number of languages in the repertoire to tolerate the otherwise problematic consonant substitutions.

Finally, the cognate pair which is perceived as the most different is *corbata* and *kravata.* Not only are they structurally very different (LD=0.50), but both consonantal and vowel changes take place in the strings. Furthermore, the onset is completely different. The low rating for the pair was expected. The Spanish cognate ratings are displayed in Table 4.

It must be noted that there was no significance found in the T-test in which the ratings of the groups of participants with and without a Romance language in the repertoire were compared. In the present study Berthele's (2011) and Vanhove's (2013) findings of correlation between the closeness of the language(s) in the repertoire and the target language and word meaning inferences has not been confirmed, at least at the level of similarity perception.

Table 4: Spanish cognates

	Mean	Variance	Overall LD	Consonantal LD
es_corbata	4,33	7,31	0,50	0,25
es_ojo	4,49	7,79	0,33	0,33
es_razón	5,05	7,59	0,40	0,20
es_huracán	6,64	7,75	0,29	0,29
es_cañón	7,05	5,90	0,33	0,33
es_bicicleta	7,26	5,79	0,44	0,22
es_cocoa	7,40	6,07	1,00	0,40
es_pistola	7,99	4,50	0,14	0,14
es_región	8,25	3,04	0,33	0,33

4.4. French

Item-related variables such as the overall and the consonantal LD, word onset and neighbourhood effect influence the ratings of the French cognates. The participants are also more sensitive to insertions and deletions as opposed to substitutions and they tend to self-pronounce certain words. The Romance language effect takes place in *poudre*, which means that the participants with one or more Romance languages in the repertoire tolerated the vowel changes in this cognate pair.

As expected, the word with the highest similarity rating is *appétit* (8.99). *Appétit* has the lowest Levenshtein distance out of all the French cognates (0.14) and there is only one consonant deletion necessary to transform it into *apetit.* Despite the change occurring in the onset, there are no other changes other than in the double consonant. Also, this cognate pair

has the lowest variance (2.54), which indicates that the participants generally considered this cognate pair to be highly similar.

The following highly-rated word is *garage*, in which the LD is low due to two changes occurring at the end of the word. It could be that the participants not only perceived it as very similar to the Croatian *garaža* because of the identical onset and low Levenshtein distance, but also given that they inferred the correspondence of the French /g/ and the Croatian /ž/ in the world.

Another cognate pair which was rated as very similar is *détail* and *detalj* (7.98). In this example there are also differences only in the final position as was the case with *garage*. Their LDs are also very similar – *détail* has LD= 0.29, whereas *garage* has a slightly higher LD (0.33). This leads to the conclusion that the nature of the differences affected the ratings, that is, *détail* is perceived as more different since the operations are a consonant insertion and a vowel deletion as opposed to the consonant substitution and vowel substitution in g*arage*.

The similarity judgement of the cognate pair *paysage* and *pejzaž* is unexpected. Firstly, the overall and the consonantal LD of the pair are very high (0.71/0.43). The only consonant in which there are not any changes is the first one, which contributed to the high result. Secondly, three consonant substitutions, a vowel substitution and deletion do take place in the rest of the word, which is why a lower rating was expected. The factor which is hypothesized to facilitate the judgement is the phonetic distance, which is much lower than the grapheme string edit distance. The conclusion is that in the case of *paysage*, participants who have judged it to be very similar to *pejzaž* successfully self-pronounced the French word.

Poudre is another example of a cognate in which only vowel changes occur. As was the case in the previous languages, it was rated relatively lower (7.09). In *poudre* and *puder* two vowel deletions and a vowel insertion take place, which was evidently problematic. However, in the T-test comparison of the groups who have and do not have a Romance language in the repertoire, a significant result occurred only in this cognate pair (p=0.04). The hypothesis is that the participants who have one or more Romance languages in the repertoires (other than French) perceived these vowel changes as more similar since they have had more input concerning the vowel structure in Romance languages and/or Latin.

Another cognate pair with low rating is *risque* and *rizik*. Even though the changes occur in word endings, the overall LD is very high (0.71) and there are three vowel operations (two

vowel deletions and a vowel insertion). As was the case in Czech and Hungarian, vowels can be the sole cause of a low similarity judgement. The similar effect takes place in *mer* and *more.* There are only vowel operations in the strings and the overall Levenshtein distance is high (0.50). Since *mer* is a monosyllabic word, the low rating can also be attributed to the abovementioned neighbourhood effect. It must be added that the cognate also has an extremely high variance (8.64), which shows the great disparity among the participants.

Lastly, by far the lowest rated cognate pair is *yaourt* and *jogurt.* The overall and the consonantal LD are lower than in some better rated cognate pairs. Yet, the changes in the entire word beginning proved to be highly problematic to most of the participants, even though it must be noted that the variance is high (8.25). Finally, no correlation was found with the number of languages. The data concerning the French cognates can be found in Table 5.

Table 5: French cognates

	Mean	Variance	Overall LD	Consonantal LD
fr_yaourt	5,40	8,25	0,43	0,29
fr_mer	5,69	8,64	0,50	0,00
fr_risque	6,84	6,76	0,71	0,29
fr_poudre	7,09	5,58	0,43	0,00
fr_paysage	7,42	4,27	0,71	0,43
fr_détail	7,98	4,79	0,29	0,14
fr_région	8,04	4,49	0,33	0,33
fr_garage	8,12	4,11	0,33	0,17
fr_appétit	8,99	2,54	0,14	0,14

4.5. German

The participants have problems with consonantal changes in German cognates; however, the cognates *Berg* and *Perücke,* where mostly vowel changes take place, are rated lower than expected. The latter implies that vowel changes are also important. Other item-related variables are also confirmed to some extent, whereas there is no correlation with the number of languages and the Romance language effect does not occur.

The rating of the cognate pair *Charakter/karakter* was the highest (8.53) despite the two consonantal changes in the word onset. However, the participants correctly identified the correspondence between the German /ch/ and Croatian /k/. As was the case with the Hungarian *csizma*, the non-existent phonetic differences and the correct inference resulted in a very high similarity judgement. It must be noted that the variance of *Charakter* is extremely low (1.75), which indicates that most of the participants inferred the phonetic correspondence.

The following two cognates are *Idee* and *Matratze*. Both cognates have changes in the word endings and the same overall LD (0.50). However, the rating of *Idee* is higher (7.92 as opposed to 6.60 for *Matratze*). The higher similarity judgement can be attributed to the lower consonantal LD as there is only one consonant insertion necessary to transform *Idee* into *ideja.* As the overall LD and the position of the changes is the same in the two cognates, the deciding factor could be the importance of consonants in similarity judgements.

In the word pair *Waage* and *vaga* the overall LD is relatively high (0.60), whereas the consonantal LD is relatively low (0.20) due to only one consonant operation. The change takes place in the first consonant, which has been proven to carry a lot of information value in the present study and in other sources (Berthele, 2011; Gooskens, Heeringa, & Beijering, 2008). Nonetheless, the similarity rating is higher than expected (6.32), which can be explained by low consonantal LD and by the fact that the word onset phonetic distance is much lower than the graphemic distance. Also, there is a phonetic correspondence between the German /w/ and the Croatian /v/ and the double vowel is only shortened by a vowel deletion. This example further proves the importance of descriptive analysis in similarity judgement analysis.

The cognate pair *Perücke* and *perika* is characterized by the identical onset, a relatively low overall LD (0.40) and the lowest consonantal LD (0.14). Nevertheless, the similarity rating is relatively low (5.31), whereas the variance is very high (9.34). Some participants evidently

find it to be very similar since they do not have issues with the vowel operations, whereas they prove to be difficult for others. However, no correlation was found for the number of languages and there was not any significance in relation to the Romance language effect. The possible issue is the difficulty of inferring a correspondence between /ck/ and /k/ when the surrounding vowels are different.

The cognate *Öl* is rated as very different from its Croatian equivalent *ulje* (4.87). The LD of the pair is the highest (0.75) and the word is monosyllabic, which means that there are more competing neighbours. However, the consonantal LD is low (0.25) and the majority operations take place in vowels. The variance of the word is extremely high (10.71), which implies that there was a great disparity among the participants. Neither the number of languages nor the Romance language effect explain the disparity. The low rating is expected, that is, it is in accordance with the starting hypotheses.

Brijeg and *Berg* form an interesting cognate pair as the differences are located in the middle of the word. What is more, the consonantal LD (0.14) is much lower than the overall LD (0.57) as there is only one consonant insertion as opposed to two vowel insertions and a vowel deletion. The vowel changes again prove very significant in the similarity judgement.

In accordance with the beginning hypotheses, the word which is perceived as the least similar to the Croatian counterpart is *Schere.* In order to convert it into *škare,* the entire onset is changed – a vowel substitution, a consonant substitution and a consonant deletion take place. The study has confirmed the tendency of the participants to rate the cognate pairs with the consonant and vowel changes in the word onset as very different from the counterparts. No relation was found of the ratings of the German cognates and the number of languages and the knowledge of Romance languages proved to have no effect either. It must be noted that most of the participants have German in their repertoire, which is the analysis of the German cognates was done on a small sample (30 participants). The similarity ratings, variance scores, as well as overall and consonantal LD values of the German cognates can be seen in Table 6.

Table 6: German cognates

	Mean	Variance	Overall LD	Consonantal LD

de_Schere	3,94	6,85	0,50	0,33
de_Berg	4,11	8,03	0,57	0,14
de_Öl	4,87	10,71	0,75	0,25
de_Perücke	5,31	9,34	0,43	0,14
de_Waage	6,32	8,08	0,60	0,20
de_Matratze	6,60	7,32	0,50	0,38
de_Idee	7,92	3,71	0,40	0,20
de_Region	7,95	4,88	0,33	0,33
de_Charakter	8,53	1,75	0,22	0,22

4.6. Swedish

In the case of Swedish cognates, item-related variables which affect the ratings are consonantal LD, type of operation, changes in the onset, neighbourhood effect. Self-pronunciation seems to have taken place in the cognates *designer,* where English may have been the supplier language and brought about high rating, and in *choklad,* which is rated much higher than expected due to false grapheme-phoneme correspondence. In these two cognates the Romance language effect also took place. It is assumed that the knowledge of Romance language(s) influenced the self-pronunciation. What is more, word ending also contributed to some ratings.

The cognate pair which is perceived as very similar is *designer* and *dizajner*. The overall LD is relatively high (0.50) and the changes take place in the first half of the word with two consonant and two vowel substitutions. It is evident that the item-related variables are not a good predictor of the similarity judgment in this word pair. However, all the participants have a certain level of English language knowledge. Since the Swedish *designer* is spelled the same as the English equivalent, it is probable that the participants drew on their knowledge of English and assessed the Swedish cognate to be very similar to the Croatian counterpart

(7.97). Moreover, in the Romance language effect T-test, the cognate *designer* proved significant (p=0.03), which implies that the participants with the knowledge of a Romance language reacted differently to the changes in the pair.

In *choklad* there are three changes necessary to transform it into *čokolada*, two vowel insertions and a consonant deletion. The pair is rated as very similar (7.75), which can be attributed to the incorrect inference of the correspondence between /ch/ and /č/. Furthermore, the vowel insertions in the middle and at the end of the word proved to be less problematic, which is in accordance with the findings which emphasize the importance of consonants. It must be noted that in this cognate pair the Romance language effect took place (p=0.01), which again indicates a difference in the perception of the participants with and without a Romance language in the repertoire.

The following two cognates with similar overall LD (0.14/0.17) display an interesting difference in similarity perception. First, *balkong* is different from *balkon* only in the final consonant. Second, *spenat* is different from *špinat* only in the first vowel. The rating of *balkong* is higher, 7.08 as opposed to 6.72 for *spenat.* The participants judged the former as more similar to its Croatian counterpart than the latter. The importance of word beginning outweighed the importance of consonants in this example.

Vin and *vino* are different only in the final vowel, the overall LD is relatively low (0.25) and there are no consonant changes. However, the similarity rating is relatively lower than in other Swedish cognates (6.62). The variable which could have contributed to the judgement is word-length, that is, the neighbourhood effect. Also, *vin* is the only Swedish cognate in which the number of languages of the participants mildly affected the rating (r= .214).

Mjölk and *ryggsäck* are low-rated cognates with relatively high overall LDs (0.63 for *mjölk* and 0.50 for *ryggsäck*). The former presupposes consonant as well as vowel changes and it was rated slightly higher than the latter (5.69). Moreover, both cognates have the highest variance in the Swedish group (7.38/8.03), which indicates that there was disparity among participants.

Finally, the lowest-rated cognate pair is *gräns* and *granica* (4.59). The onsets of the words are identical, but there are two vowel insertions and a consonant substitution in the word ending. The different word ending and the nature of operations (insertions) are problematic to the

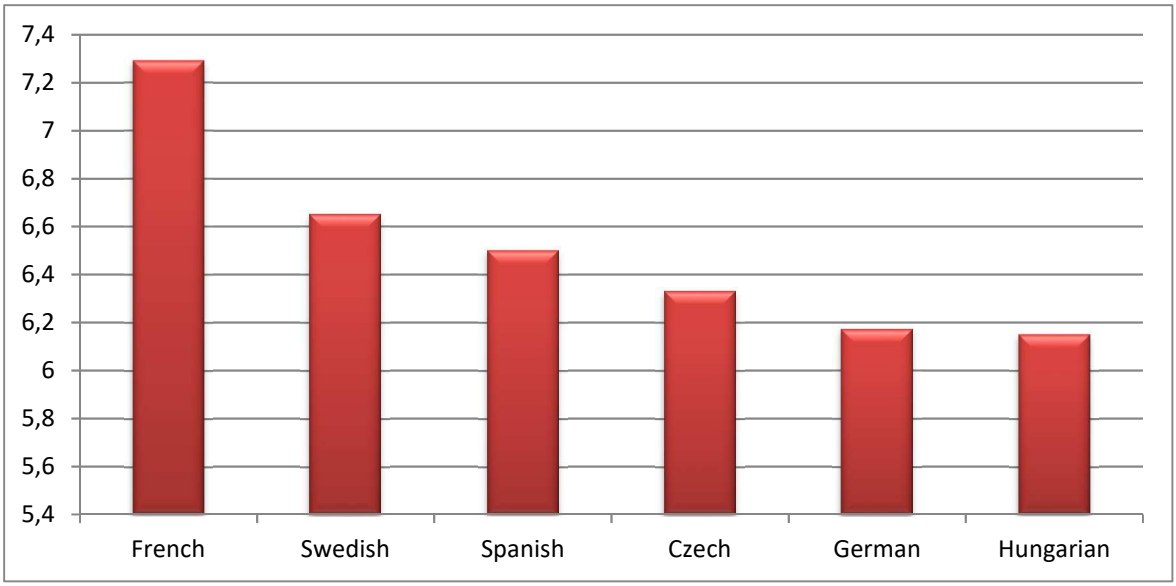
participants in the case of this word pair. In Table 7 the data on all Swedish cognates can be found.

Table 7: Swedish cognates

	Mean	Variance	Overall LD	Consonantal LD
sw_gräns	4,59	6,51	0,43	0,14
sw_ryggsäck	5,57	8,03	0,50	0,50
sw_mjölk	5,69	7,38	0,63	0,25
sw_vin	6,62	7,06	0,25	0,00
sw_spenat	6,72	4,56	0,17	0,00
sw_balkong	7,08	5,49	0,14	0,14
sw_choklad	7,75	4,12	0,33	0,11
sw_region	7,88	4,49	0,33	0,33
sw_designer	7,97	4,36	0,50	0,25

4.7. Psychotypology at the language system level

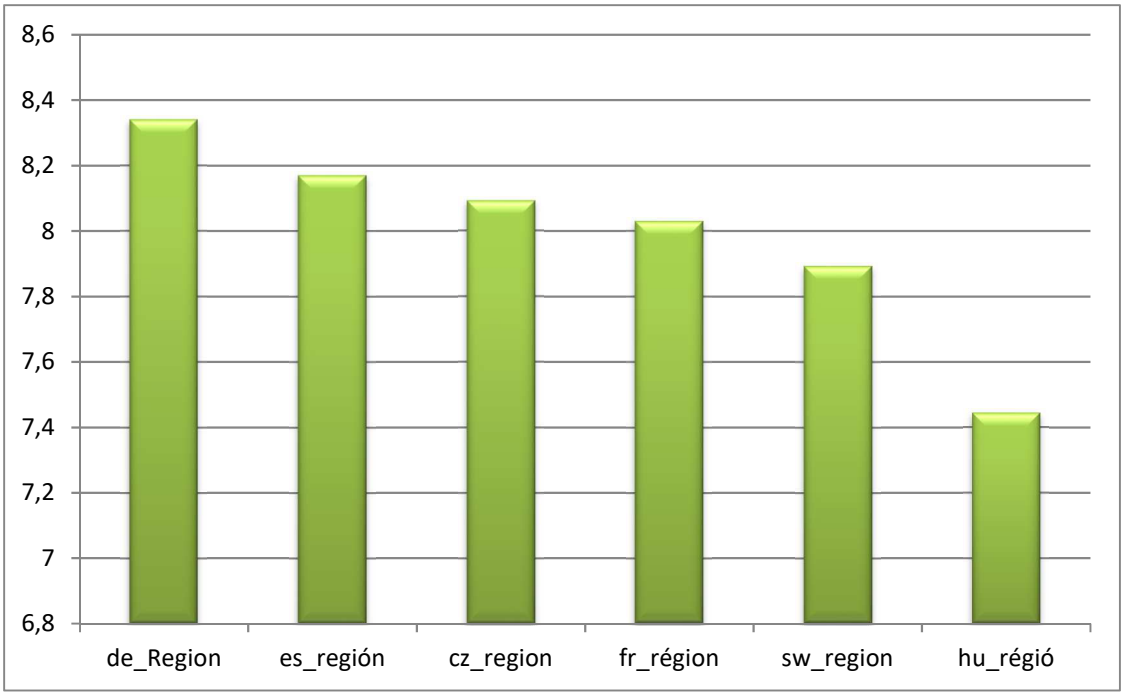
It can be assumed that the mean ratings of all the cognates in each language can display the participants' perception of the six languages. Again, the hypotheses are that Czech would be rated higher than Hungarian, Spanish higher than French and German higher than Swedish. There are different criteria which can be adopted in such assumptions and a certain degree of subjectivity. However, it can be asserted that due to the typological distance, Hungarian would be rated as very different, which can be also expected for Swedish due to the participants' lack of contact and experience with the language. Graph 1 contains the mean ratings for each language.



Graph 1: Mean similarity ratings for each language in the study

As can be seen from the results, French and Swedish are rated as the most similar to Croatian, which does not correspond to the expectations. It is also surprising that Czech has such a low rating given that it is the only language in the study which is in the same family as Croatian. It must be noted that Hungarian is the lowest rated language. However, the reliability of the data is highly questionable since the item-related variables are not equally distributed or controlled for across all six languages. For example, the mean Levenshtein distance value for all German cognates is 0.48 and 0.31 for the Hungarian cognates, which means that the German cognates were formally less similar to the Croatian equivalents than Hungarian.

Nonetheless, the loanword *region* enables a much more reliable analysis. Even though it is the only such example, it has equal overall LD (0.33) across all the languages in the study and the vowel and consonant changes take place in the final part of the word, which means that the item-related variables are controlled for. Graph 2 shows the mean ratings for all instance of the loanword.



Graph 2: Mean ratings for the loanword *region*

The results correspond to the hypotheses about the Croatian participants' perception of languages. The loanword is among the top three highest rated cognates in each language, which is in accordance with the claim by Kurschner et al. (2008) that loanwords are easier to understand than native cognates as they have not been integrated into the target language to the same extent.

The highest-rated word in Czech is precisely *region.* Not only is it perceived as the most similar, but it has the lowest variance of all cognates (4.22). It appears that most of the participants were unanimous in their high assessment, even though the operations necessary to transform *region* into *regija* include vowel-consonant and consonant-vowel substitutions. Item-related factors which could contribute to the high assessment are the relatively low Levenshtein distance and the lack of differences in the onset.

The French *région* was also perceived as very similar to the Croatian counterpart (8.04). Even though it is not the highest rated cognate, the rating of the French loanword resembles the similarity judgement of its Spanish and Czech equivalent. Aside from the psychotypology at the language system level, the lower position on the similarity rating scale could be attributed to the differences between the groups of cognates in general. The surrounding word pairs affected the similarity judgement of the participants as well as numerous other variables.

What is more, the lowest rated languages are Swedish and Hungarian. In both tests Hungarian is the lowest rated language, which indicates that formal distance in typologically very distant languages plays an important role in psychotypology. The results of the mean ratings of the internationalism are in accordance with expectations. Since the item-related variables are controlled for in all six languages, it can be stated that the ratings of *region* reflect the perception of the six language systems of the Croatian participants. However, it must be acknowledged that it is only one word and that numerous other variables could have affected the ratings. More on the limitations of the study can be found in the following subsection.

5. Limitations

The present study is an ambitious attempt at tackling the issue of language perception. Language typology is an excellent and necessary way to gain insight into how similar or different languages are. However, as it is often the case with any matter that involves human cognition and perception, human beings do not perceive the objective distance between languages in the same way. Approaching this issue empirically is a difficult task. The variables which are at play in the process of cognate similarity judgements are numerous, particularly if one considers the number of supplier languages and the interactions between item-related variables of the cognates in all the languages in the participants' repertoires. Therefore, Vanhove and Berthele (2015) insist that "researchers consequently need to take rather arbitrary decisions about which variables to include with respect to which potential supplier languages so that the set of predictors remain of a manageable size" (p.2).

Even though the sample was intended to be heterogeneous, it can be said that the heterogeneity of the participants and the items is the biggest obstacle to making any empirical, concrete inferences. What is more, the number of variables in the study is large, yet the number of variables which are not controlled for is even bigger. For example, there is not a systematic account of all the potential supplier languages and even the surrounding words could have affected the ratings. The complexity of the methodology is evident and it is suggested that more focus is directed to one or two variables in future studies.

6. Conclusion

One could claim that the present research discovered nothing, yet uncovered a lot. The discrepancy between the formal, objective language distance and the perception of that distance has attracted interest of various authors (Berthele, 2012; Vanhove et al., 2013; Möller et al., 2015; Gooskens, 2007; Kaivapalu et al., 2017 and many more). Most of the studies tackle cognate recognition and intelligibility, which form an integral part of receptive multilingualism. The present study is focused on the factors which affect the perception of cognate similarity as it is claimed that similarity perception underlies the process of cognate recognition. When approaching an unknown language, language learners or users rely on what they believe to be similar to the resources they already possess. Contact with a new language can be described as a metaphorical battlefield where more or less similar cognates are the front-line troops. However, in order to be able to conquer the front line, one has to notice it and have the adequate strategies and tactics to make use of it.

The factors which influence the perception of what is similar and useful in an unknown language are abundant. On one side of the battlefield there is the language learner or user who is moulded by the biological, social, mental and situational characteristics. The previous contact and experience with different languages, dialects and varieties can be seen as the artillery. The language user needs to perceive the connection between the familiar and unfamiliar in order to optimally use the "weapon". The unfamiliar language and all its linguistic determinants are on the other side of the battlefield. On the macro level, the perception of the opponent also determines the way one will approach fighting the battle.

The battlefield metaphor illustrates the complexity of variables which play a role in the similarity perception, especially in a study with two base languages, six target languages, 54 cognate pairs and different language constellations of the participants. The ratings of the participants were affected by the item-related variables. However, not all item-related variables are equally important in every language and cognate pair, and they also interact with participant-related variables. For example, the Spanish cognate *razón* is rated lower than expected (5.05) despite the identical onset, whereas a difference in the onset affects the low rating of *hluma* in Czech. What is more, the variance of *razón* is high, which indicates that there was a disparity among the participants, that is, some participants were more sensitive to the differences in the word ending than others. The participants also made assumptions about

phoneme-grapheme correspondences, which shows that they engaged in “dynamic construction of hypothetical grammars” (Berthele, 2011).

The question of which participants are more affected by which changes in the items and what role the overall perception of a language system plays in the process remains open, even though certain tendencies are revealed. The word onset proved to be very important, as well as consonant and vowel changes. Furthermore, phonetic distance must not be disregarded as well as the supplier languages and dialects in the participants’ repertoires. It also must be added that the similarity perception is not a state but a dynamic process in which a change in one variable can cause a ripple effect in the others. However, discovering the interplay of factors which determine whether and how the language learner or user will use their “weapons” merits further, in-depth research.

References

Ashby, M. & Maidment, J. (2005). *Introducing Phonetic Science (Cambridge Introductions to Language and Linguistics).* Cambridge: Cambridge University Press. doi:10.1017/CBO9780511808852

Beijering, K. & Gooskens, C. & Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands,* 25(1). 25. 10.1075/avt.25.05bei.

Berthele, R. (2008). Dialekt-Standard Situationen als embryonale Mehrsprachigkeit. Erkenntnisse zum interlingualen Potenzial des Provinzlerdaseins. In: Mattheier, K.J. & Lenz, A. (Eds.) *Dialektsoziologie/Dialect Sociology/Sociologie du Dialecte. Sociolinguistica*, 22. Tübingen: Niemeyer, pp. 87-107.

Berthele, R. (2011). On abduction in receptive multilingualism. Evidence from cognate guessing tasks. *Applied Linguistics Review*, 2, pp. 191-220. doi:10.1515/9783110239331.191

Broerse, A. C., & Zwaan, E. J. (1966). The information value of initial letters in the identification of words. *Journal of Verbal Learning & Verbal Behavior*, 5(5), pp. 441-446. http://dx.doi.org/10.1016/S0022-5371(66)80058-0

Gooskens, C. (2007). The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. *Journal of Multilingualism and Multicultural Development,* 28(6), pp. 445-467. 10.2167/jmmd511.0

Gooskens, C. & Heeringa, W. & Beijering, K. (2008). Phonetic and Lexical Predictors of Intelligibility. *International Journal of Humanities and Arts Computing*, 2, pp. 63-81. 10.3366/E1753854809000317.

Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic Influence in Language and Cognition.* Routledge.

Kaivapalu, A., & Martin, M. (2017). Perceived similarity between written Estonian and Finnish : Strings of letters or morphological units?. *Nordic Journal of Linguistics*, 40 (2), 149-174. doi:10.1017/s0332586517000142

Kellerman, E. (1983). Now You See It, Now You Don't. In: Gass, S. & Selinker, L. (Eds.), *Language Transfer in Language Learning*. Rowley, MA: Newbury House, pp. 112-134.

Kürschner, S. & Gooskens, C. & van Bezooijen, R. (2008). Linguistic Determinants of the Intelligibility of Swedish Words among Danes. *International Journal of Humanities and Arts Computing*, 2, pp. 83-100. 10.3366/E1753854809000329.

Letica Krevelj, Stela (2014). Crosslinguistic interaction in acquiring English as L3: role of psychotypology and L2 status. Unpublished doctoral thesis. University of Zagreb, Faculty of Humanities and Social Sciences. Zagreb, Croatia.

Letica Krevelj, S. (2016). Multilinguals' perceptions of crosslinguistic similarity and relative ease of learning genealogically unrelated languages. *Studia Romanica et Anglica Zagrabiensia*, 61 (-), 175-205. https://hrcak.srce.hr/180516

Meissner, F.J. (1997). Philologiestudenten lesen in fremden romanischen Sprachen. Konsequenzen für die Mehrsprachigkeitsdidaktik aus einem empirischen Vergleich. In: Meissner, F.J. (Ed.), *Interaktiver Fremdsprachenunterricht. Wege zu authentischer Kommunikation. Ludger Schiffler zum 60. Geburtstag*. Tübingen: Narr, pp. 25-44.

Möller, Robert & Zeevaert, Ludger. (2015). Investigating word recognition in intercomprehension: Methods and findings. *Linguistics*. 53(2). 10.1515/ling-2015-0006.

Otwinowska-Kasztelanic, Agnieszka. (2011). Awareness and affordances: Multilinguals versus bilinguals and their perceptions of cognates. In: De Angelis, G. and Dewaele J-M (Eds.) (2011) *New trends in crosslinguistic influence and multilingualism research*. Multilingual Matters, pp. 1-18. 10.21832/9781847694430-002.

Peyer, Elisabeth & Kaiser, Irmtraud & Berthele, Raphael. (2010). The multilingual reader: Advantages in understanding and decoding German sentence structure when reading German as an L3. *International Journal of Multilingualism*,7(3). 7. 10.1080/14790711003599443.

Ringbom, H. (2006). *Cross-linguistic Similarity in Foreign Language Learning.* Multilingual Matters.

Schmitt, N. (1997). Vocabulary learning strategies. In: Schmitt, N. and McCarthy, M. (Eds.) (1997) *Vocabulary. Description, Acquisition and Pedagogy*. Cambridge: CUP, pp. 199-227

Utgof, D. (2008). The Perception of Lexical Similarities Between L2 English and L3 Swedish (Lingköping University). https://www.researchgate.net/publication/279503662

Vanhove, J. & Berthele, R. (2015). Item-related determinants of cognate guessing in multilinguals. In: De Angelis, G., Jessner, U. & Kresić, M. (Eds.) *Crosslinguistic influence and crosslinguistic interaction in multilingual language learning*. London: Bloomsbury, pp. 95-118.

Vanhove, J. & Berthele, R. (2017). Interactions between formal distance and participant-related variables in receptive multilingualism. *IRAL - International Review of Applied Linguistics in Language Teaching,* 55(1). 10.1515/iral-2017-0007.

Zipf, George K. (1949). *Human behaviour and the principle of least effort.* Addison-Wesley Press.

Appendix – Questionnaire

Poštovani,

Zahvaljujem Vam na pristanku na sudjelovanje u ovom istraživanju. Podaci prikupljeni ovim upitnikom koristit će se isključivo u istraživačke svrhe. Upitnik je anoniman, a rezultati će se prikazati samo u kumulativnom obliku.

UPITNIK O SLIČNOSTI IZMEĐU PAROVA RIJEČI

I. Biografski podaci (molim Vas da nadopunite ili zaokružite):

1. Dob:_____

   Spol: M / Ž

   Mjesto stanovanja: _____

   Mjesto rođenja: _____

   Razina obrazovanja:

   a) osnovna škola

   b) srednja škola

   c) fakultet

   Materinski jezik: _____

   Govorite li neki dijalekt? DA / NE. Ako da, navedite koji: _____

2. Navedite sve jezike koje ste učili ili kojima ste na neki način bili izloženi, te na skali označite svoju procjenu znanja svakog od jezika.

Jezik 1 _____

Znanje: početničko 1 2 3 4 5 napredno

Jezik 2 _____

Znanje: početničko 1 2 3 4 5 napredno

Jezik 3 _____

Znanje: početničko 1 2 3 4 5 napredno

Jezik 4 _____

Znanje: početničko 1 2 3 4 5 napredno

Jezik 5 _____

Znanje: početničko 1 2 3 4 5 napredno

Jezik 6 _____

Znanje: početničko 1 2 3 4 5 napredno

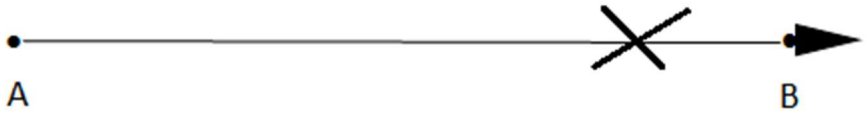
3. Studirate li jezike ili se bavite nekom strukom koja je povezana s jezicima? DA / NE

Ako da, molim Vas da navedete predmet studija ili struku:

_____

II. S lijeve strane pravca navedena je riječ na stranom jeziku, a s desne strane njen prijevod na hrvatskom jeziku. Kako biste procijenili sličnost tih parova riječi na pravcu između točke A i B, ako točka A označava da su te riječi potpuno različite, a točka B da su potpuno iste?

Molimo da Vašu procjenu sličnosti riječi naznačite na pravcu <u>pomoću križića</u> kao što je prikazano na primjeru:

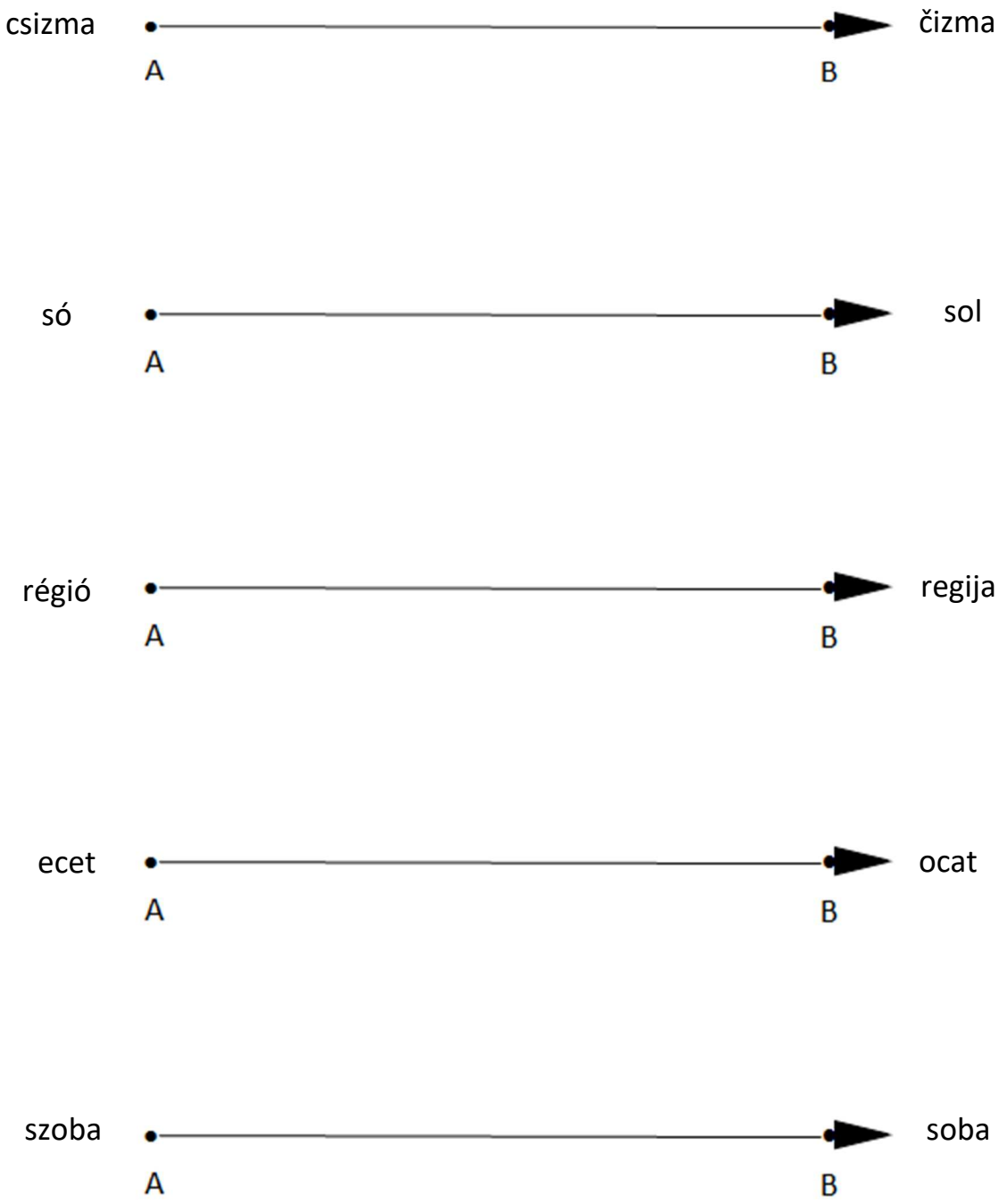


1. **Pokušajte procijeniti sličnost sljedećih parova riječi:**

2. **Molim Vas da sve ostale zadatke riješite na isti način. Riječi s lijeve strane će biti na različitim jezicima, a na početku svakog dijela navedeno je o kojem se jeziku radi.**

MAĐARSKI

točka A = potpuno različite
točka B = potpuno iste




vacsora
večera
A
B

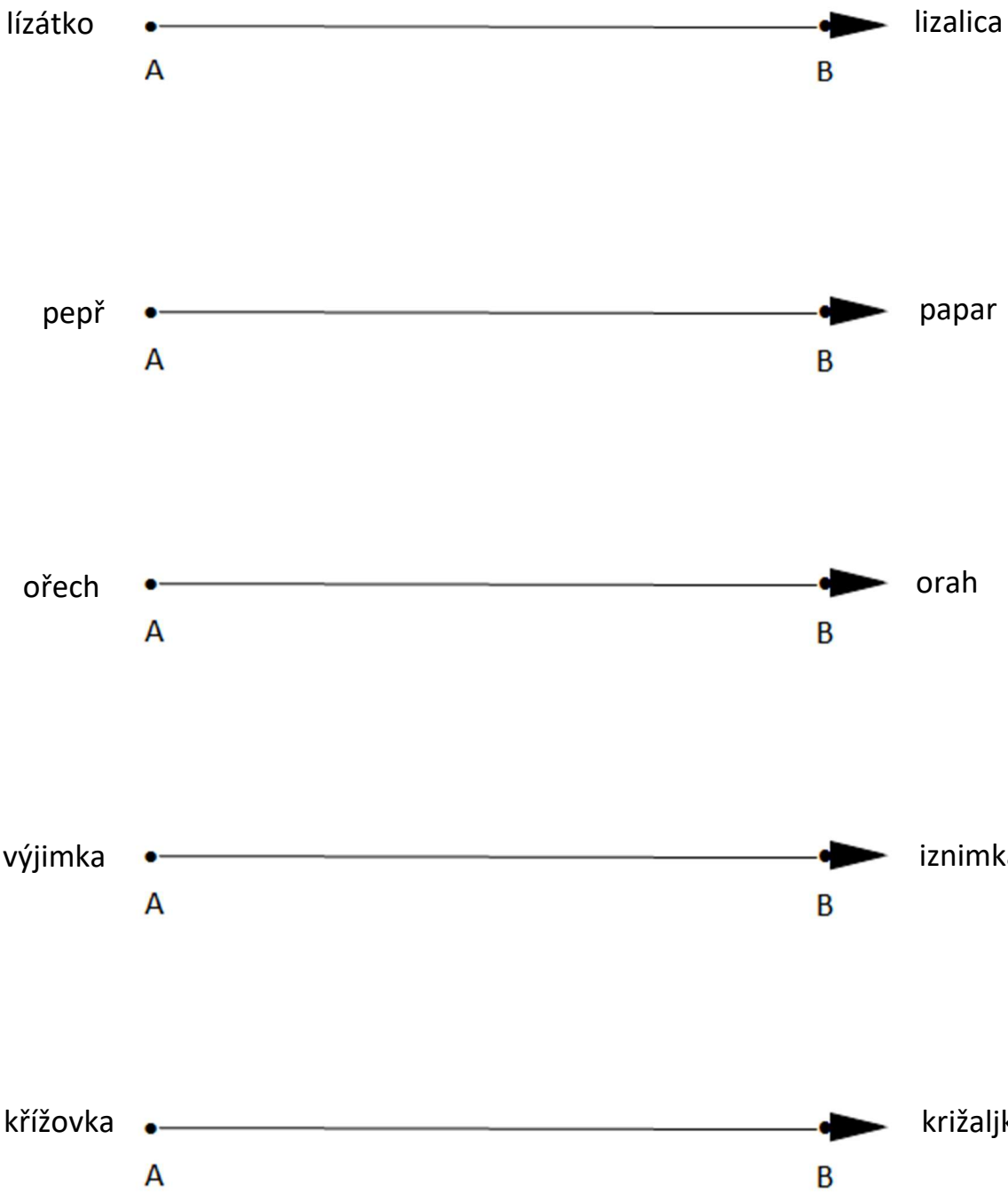

görcs
grč
A
B


palacsinta
palačinka
A
B


kóró
korov
A
B

ČEŠKI

točka A = potpuno različite
točka B = potpuno iste




hluma
gluma
A
B

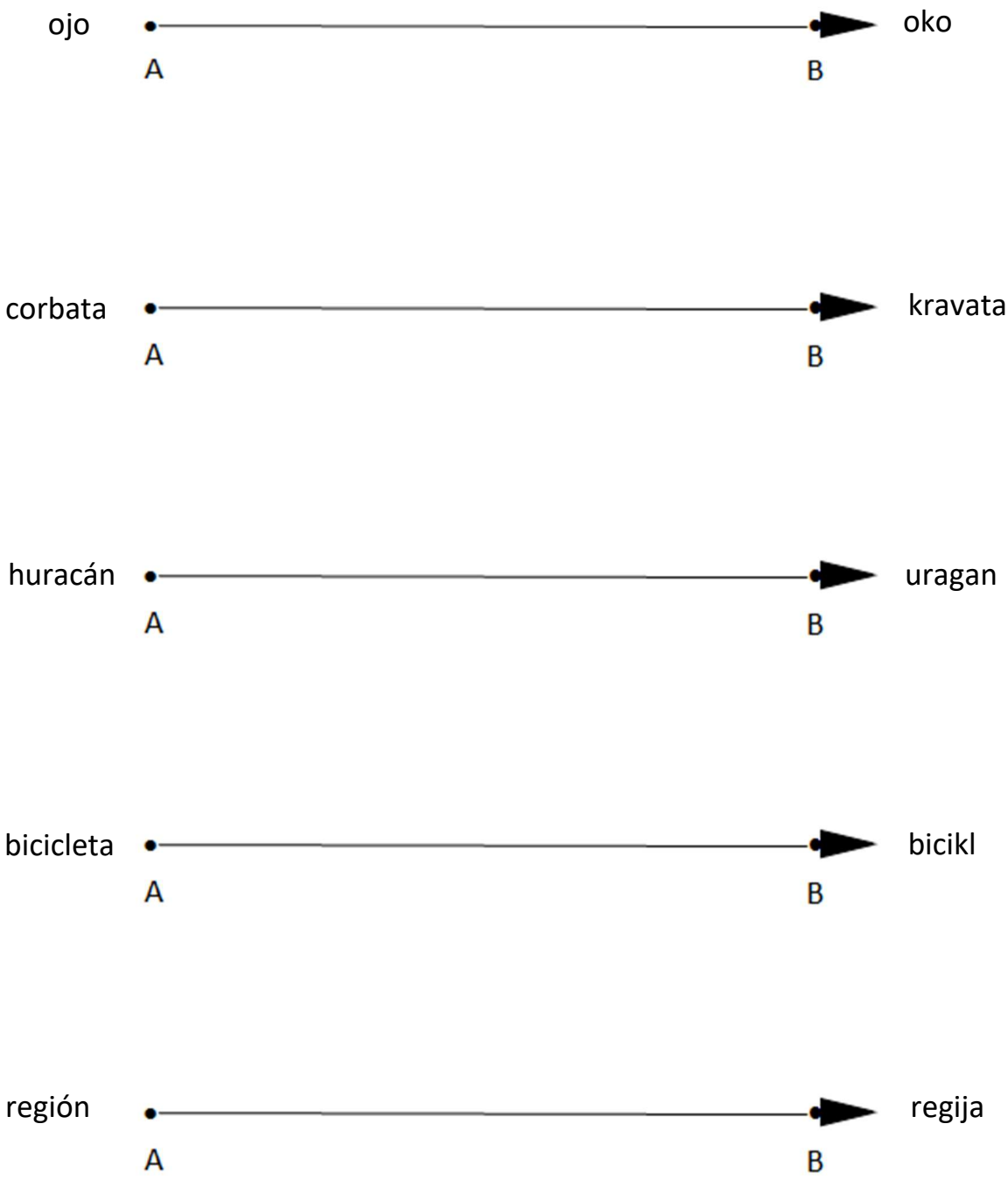

region
regija
A
B


důkaz
dokaz
A
B


ucho
uho
A
B

ŠPANJOLSKI

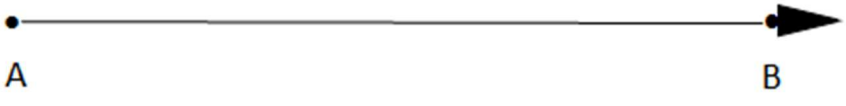
točka A = potpuno različite
točka B = potpuno iste



cañón  kanjon

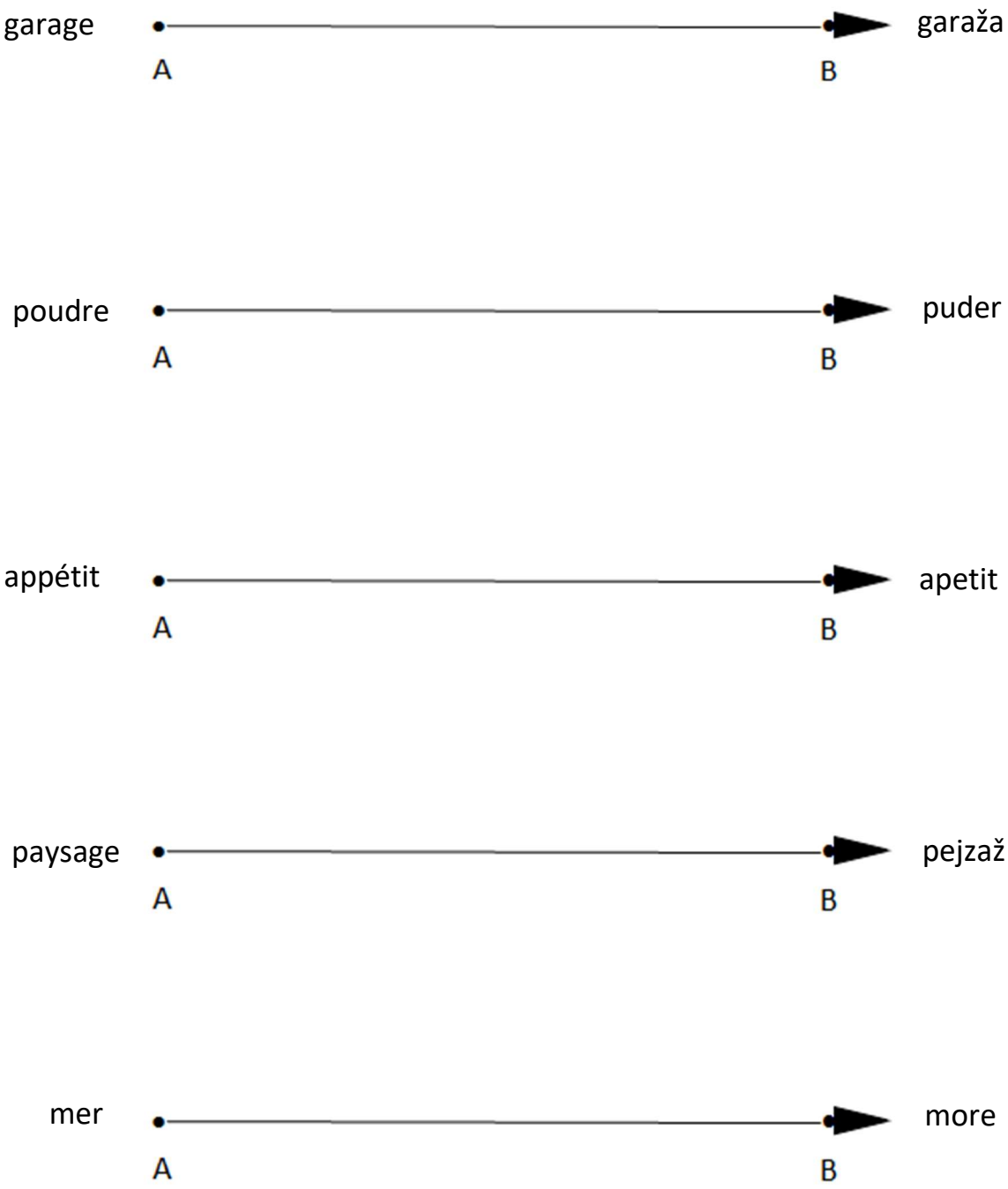
cocoa  kakao

pistola  pištolj

razón  razum

FRAN□USKI

točka A = potpuno različite
točka B = potpuno iste




yaourt
jogurt
A
B

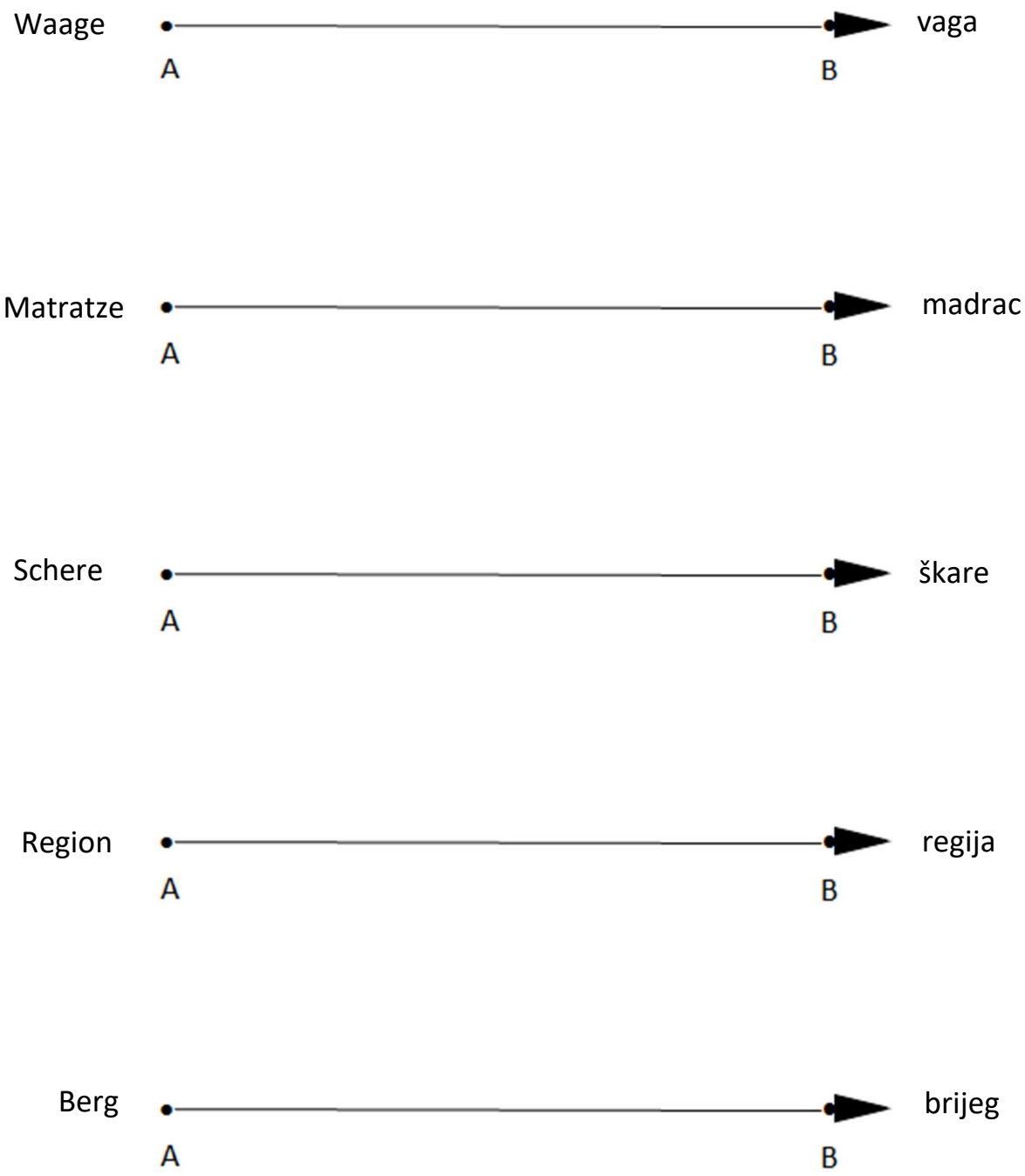

risque
rizik
A
B

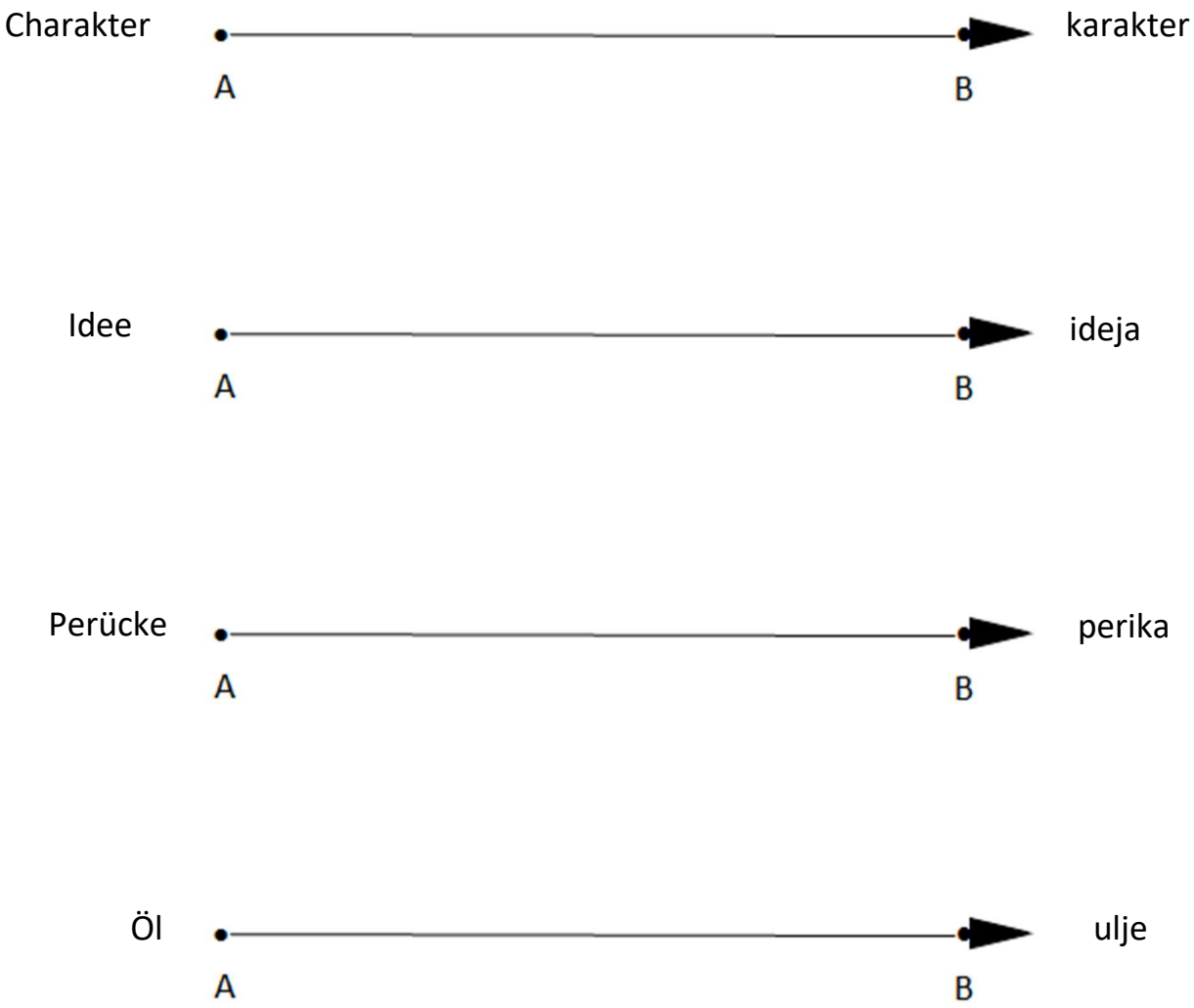

détail
detalj
A
B


région
regija
A
B

NJEMAČKI

točka A = potpuno različite
točka B = potpuno iste

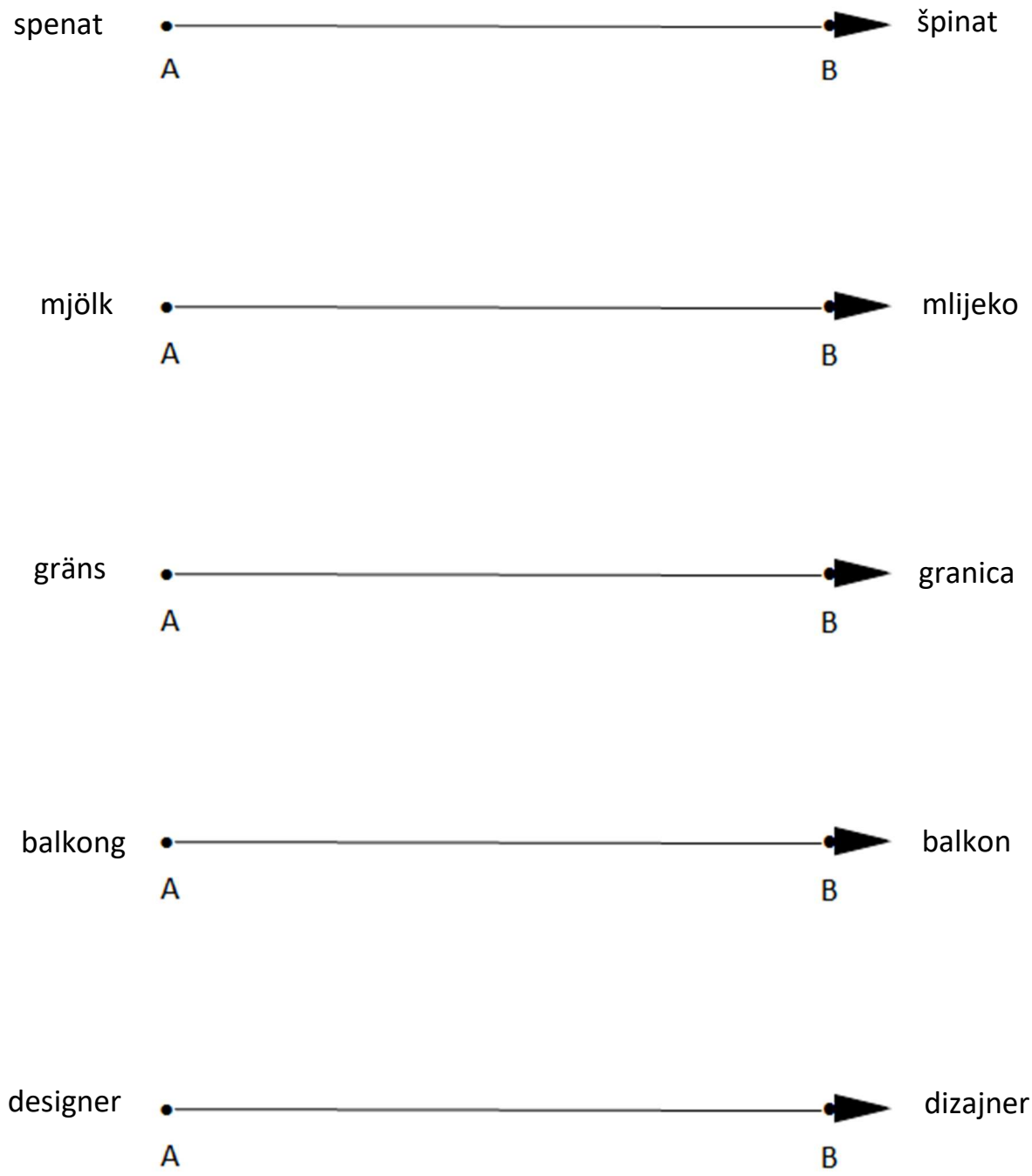



Charakter
karakter
A
B
Idee
ideja
A
B
Perücke
perika
A
B
Öl
ulje
A
B

ŠVEDSKI

točka A = potpuno različite

točka B = potpuno iste



region

A B

regija

ryggsäck

A B

ruksak

vin

A B

vino

choklad

A B

čokolada